Received: 20 June 2019 | Accepted: 2 January 2020

DOI: 10.1002/pros.23952

ORIGINAL ARTICLE

The Prostate WILEY

A qualitative transcriptional signature for predicting the biochemical recurrence risk of prostate cancer patients after radical prostatectomy

Xiang Li[1,2,3] | Haiyan Huang[1] | Jiahui Zhang[1] | Fengle Jiang[1] | Yating Guo[1] | Yidan Shi[1] | Zheng Guo PhD[1,2,3] | Lu Ao PhD[1,2,3]

[1]Department of Bioinformatics, Key Laboratory of Ministry of Education for Gastrointestinal Cancer, The School of Basic Medical Sciences, Fujian Medical University, Fuzhou, China

[2]Key Laboratory of Medical Bioinformatics, Fujian Medical University, Fuzhou, China

[3]Fujian Key Laboratory of Tumor Microbiology, Fujian Medical University, Fuzhou, China

Correspondence
Lu Ao and Zheng Guo, Department of Bioinformatics, Key Laboratory of Ministry of Education for Gastrointestinal Cancer, The School of Basic Medical Sciences, Fujian Medical University, Fuzhou 350122, China.
Email: lukey@fjmu.edu.cn (LA) and guoz@ems.hrbmu.edu.cn (ZG)

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 81602738, 81872396; Young and Middle-aged Backbone Training Project in the Health System of Fujian Province, Grant/Award Number: 2017-ZQN-56; Outstanding Youth Scientific Research Personnel Training Program in Fujian Province University, Grant/Award Number: 2017B018; Joint Scientific and Technology Innovation Fund of Fujian Province, Grant/Award Number: 2018Y9065

Abstract

Background: The qualitative transcriptional characteristics, the within-sample relative expression orderings (REOs) of genes, are highly robust against batch effects and sample quality variations. Hence, we develop a qualitative transcriptional signature based on REOs to predict the biochemical recurrence risk of prostate cancer (PCa) patients after radical prostatectomy.

Methods: Gene pairs with REOs significantly correlated with the biochemical recurrence-free survival (BFS) were identified from 131 PCa samples in the training data set. From these gene pairs, we selected a qualitative transcriptional signature based on the within-sample REOs of gene pairs which could predict the recurrence risk of PCa patients after radical prostatectomy.

Results: A signature consisting of 74 gene pairs, named 74-GPS, was developed for predicting the recurrence risk of PCa patients after radical prostatectomy based on the majority voting rule that a sample was assigned as high risk when at least 37 gene pairs of the 74-GPS voted for high risk; otherwise, low risk. The signature was validated in six independent datasets produced by different platforms. In each of the validation datasets, the Kaplan-Meier survival analysis showed that the average BFS of the low-risk group was significantly better than that of the high-risk group. Analyses of multiomics data of PCa samples from TCGA suggested that both the epigenomic and genomic alternations could cause the reproducible transcriptional differences between the two different prognostic groups.

Conclusions: The proposed qualitative transcriptional signature can robustly stratify PCa patients after radical prostatectomy into two groups with different recurrence risk and distinct multiomics characteristics. Hence, 74-GPS may serve as a helpful tool for guiding the management of PCa patients with radical prostatectomy at the individual level.

KEYWORDS

biochemical recurrence-free survival, prostate cancer, qualitative signature, relative expression orderings

1 | INTRODUCTION

Prostate cancer (PCa) is the second most frequently diagnosed malignant tumor in men worldwide, with the highest morbidity rates in the developed countries.[1,2] In China, PCa has the most rapid rise of incidence along with the increases of the aging population and the implementation of advanced detection services in recent decades.[3] The standard method to treat localized PCa patients is radical prostatectomy, while approximately 20% to 40% of patients will suffer from biochemical recurrence in 10 years.[4,5] The prostate-specific antigen (PSA) level is an important indicator of biochemical recurrence for localized and locally advanced PCa after radical prostatectomy.[6] Nevertheless, some PCa patients with poor prognoses have a low PSA level.[7,8] The currently available clinical-pathological features, such as the Gleason grade group, clinical and pathological stage and surgical margin,[9,10] are unable to provide accurate predictions for biochemical recurrence.[11–13] Thus, it is crucial to develop an accurate prognostic signature to predict the recurrence risk for PCa patients after radical prostatectomy.

High-throughput microarray and RNA-sequencing technologies facilitate researchers developing transcriptional prognostic signatures for PCa patients.[14–16] However, most of the reported transcriptional signatures depend on risk threshold values summarized from the quantitative expression measurements of the signature genes,[14–16] which are easily vulnerable to the measurement variations from batch effects introduced by laboratory conditions, reagent lots, and personal differences. In fact, subtle quantitative values of gene expression measurements are quite error-prone.[17] Data normalization methods for removing batch effects might even exacerbate the batch problems[18,19] and these methods are not suitable for individualized analysis of clinical application. On the contrary, the within-sample relative expression orderings (REOs) of genes that are the qualitative features of transcription have been proved to be robust against experimental batch effects and differences in probe designs of different platforms.[20,21] The within-sample REO is a promising feature for building robust classifiers, for example top-scoring pair (TSP)[22] and k-TSP[23] with existing R packages.[24,25] Besides, the within-sample REO can be robustly analyzed individual sample without normalization, which is suitable for individualized application in clinical practice.

More importantly, our previous studies have demonstrated that different from the signatures based on quantitative expression measurements of the signature genes, the REOs-based qualitative signatures are rather insensitive to the tumor cell percentage difference of specimen sampled from different parts of the same tumor,[26] the inescapably partial RNA degradation[27,28] and amplification bias of low-input RNA samples.[29] Based on these unique advantages of the within-sample REOs, we have developed the qualitative REOs-based signatures for predicting the prognosis of breast cancer,[30,31] colorectal cancer,[32] gastric cancer,[33] liver cancer,[34] and lung cancer.[35] Thus, it is worthwhile to develop a qualitative prognostic signature for PCa patients after radical prostatectomy.

In this study, a qualitative REOs-based signature consisting of 74 gene pairs, named as 74-GPS, was developed to predict the recurrence risk of PCa patients using 131 samples in the training data set. A sample was assigned as high-risk when at least 37 gene pairs of the 74-GPS voted for high risk; otherwise, low risk. This signature was validated in six independent datasets produced by different platforms, totally including 660 fresh-frozen (FF) samples and 106 formalin-fixed paraffin-embedded (FFPE) samples. Using the multiomics data of PCa samples from The Cancer Genome Atlas (TCGA), we analyzed the distinct transcriptomic, epigenetic, and genomic differences between the two prognostic groups. The results might be helpful for understanding the mechanisms of different prognoses and guiding the management for PCa patients.

2 | MATERIALS AND METHODS

2.1 | Data collection and data preprocessing

Data for PCa were downloaded from the Gene Expression Omnibus[36] (GEO, http://www.ncbi.nlm.nih.gov/geo/) and TCGA (http://cancergenome.nih.gov/) database. A total of 791 FF samples from six datasets and 106 FFPE samples from the GSE54460 data set with BFS data were analyzed, as described in Table 1. These datasets were measured by different platforms, including single-channel microarray, dual-channel microarray, and next-generation sequencing (NGS). For datasets measured by the Affymetrix platform, a robust multi-array average (RMA) algorithm was used to process the raw mRNA expression data (.CEL files).[37] For datasets measured by the Illumina platform and the Stanford Functional Genomics Facility dual-channel platform, the processed data were directly used.

The level 3 mRNA-seq profiles, DNA methylation profiles, and copy number profiles used in the study were obtained from cBioPortal (http://www.cbioportal.org/),[38,39] which the gene expression values, the methylation values (β) of the CpG sites or the copy number alternation status of regions had already been mapped to Entrez gene IDs. We directly downloaded these processed data. Overall, 20 436 genes for gene expression data, 16 182 genes for DNA methylation data and 23 286 genes for copy number data were analyzed in this study, respectively. The level 2 gene mutation data were downloaded from TCGA portal. A discrete mutation profile including 11 249 genes only with the nonsynonymous mutations were generated.

2.2 | Survival analysis

The BFS time was calculated from the date of the resection to the date of biochemical recurrence or the date of the last follow-up visit. The Cox proportional hazards regression model was used to calculate the hazard ratios (HRs) and corresponding 95% confidence intervals (CIs) and estimate the independent prognostic significance of the signature after adjustment for clinic pathological factors including Gleason grade group, surgical margins, preoperative PSA and pathological stage. Harrell's concordance index (C-index)[40] was used to quantify the overall concordance between the predicted risk

TABLE 1 Description of the datasets used in this study

	PC131	PC332	PC111	PC92	PC89	PC36	PC106
Accession	GSE21032	TCGA	GSE70768	GSE70769	GSE40272	GSE46602	GSE54460
Platform	GPL10264	Illumina Hiseq-RNAseqV2	GPL10558	GPL10558	GPL9497	GPL570	GPL11154
Sample size	131	332	111	92	89	36	106
Sample type	FF	FF	FF	FF	FF	FF	FFPE
Age	58	61	62	-	62	63	61.7
Median follow-up period (mo)	54.5 (1.9-149.2)	28.9 (0.2-163.6)	34 (2-66.8)	79.6 (1.8-122.3)	43.3 (0-116)	-	68.5 (0.7-180.6)
Pathologic stage							
T1-T2	85	-	33	48	-	19	87
T3-T4	46	-	78	42	-	17	18
NA	0	-	0	2	-	0	1
Median preoperative PSA (ng/mL)	8.5 (1.1-132)	9.8 (0.8-87)	8.6 (3.2-23.7)	11 (1.5-117)	6.7 (2.1-44.5)	18.2 (5.3-42.5)	10.9 (1.8-72.6)
Gleason grade group							
1	41	-	17	20	13	16	11
2-3	74	-	85	55	65	15	80
4	8	-	8	5	4	4	10
5	7	-	1	10	7	1	5
NA	1	-	0	2	0	0	0
Surgical margin							
positive	31	69	26	42	10	16	40
negative	100	244	85	50	78	20	61
NA	0	19	0	0	1	0	5

Abbreviations: FF, fresh-frozen; FFPE, formalin-fixed paraffin-embedded; NA, not available; PSA, prostate-specific antigen.

classification and the BFS time. The log-rank tests[41] was used to compute the *p*-value for the differences between the Kaplan-Meier survival curves of BFS in two distinct subgroups.

2.3 | Development of the qualitative signature

For a gene pair (G_a, G_b), gene a and gene b with expression levels of E_a and E_b, its REO ($E_a>E_b$ or $E_a<E_b$) classified all samples into two subgroups. If the two subgroups had significantly different BFS by the univariate Cox proportional-hazards regression model, the gene pair was defined as a prognosis-associated gene pair. The Storey procedure was used to adjust the *P*-values into false discovery rate (FDR).[42] The significant level was set at 20%. All prognosis-associated gene pairs were sorted in descending order according to their C-index values. A forward selection procedure was applied to find the best subset of the prognosis-related gene pairs that achieved the highest C-index in the training data set. The first gene pair with the highest C-index was selected as a seed and the other prognosis-related gene pairs were added into the seed one by one based on the descending C-index value if the gene pair can improve the C-index. The subset of prognosis-related gene pairs with the highest C-index was chosen as the final prognostic signature. The voting rule was as follows: a patient was classified into the high-risk group when at least 50% of the gene pairs voting for high risk; otherwise, the patient was classified into the low-risk group. The R-codes for developing the REOs-based signature were available in Supporting Information Methods.

2.4 | Analysis of epigenomic and genomic data

The RankCompV2 method, which is insensitive to batch effects,[43] was used to identify differentially expressed genes (DEGs) between the high-risk and low-risk groups of PCa samples. the Wilcoxon rank-sum test was used to identify differentially methylated genes (DMGs). Fisher's exact test was used to detect genes whose frequencies of copy number alteration or mutation were significantly different between two prognostic groups of TCGA samples.

2.5 | Direction concordance scores

If *k* genes were overlapped between two DEGs lists identified from two datasets, of which *s* genes showed the same dysregulated direction (both up- or downregulated in the high-risk group compared to the low-risk group), then the direction concordance score was computed as *s/k*. Similarly, if *k* genes were overlapped between two DMGs lists identified from two datasets, of which *s* genes showed both hyper- or hypomethylated in the high-risk group compared with the low-risk group, then the direction concordance score was computed as *s/k*. For *k* DMGs, if *s* genes were upregulated (or downregulated) and

correspondingly hypomethylated (or hypermethylated), the direction concordance score was computed as *s*/*k*. This score was used to calculate the reproducibility of DEGs identified from multiple independent datasets and the consistency between DEGs and DMGs. The cumulative binomial distribution model[44] was used to evaluate whether the case of observing a direction concordance score of *s*/*k* is random:

$$p = 1 - \sum_{i=0}^{s-1} \binom{k}{i} p_e^i (1 - p_e)^{k-i}$$

where $p_e = 0.5$ is the probability of one gene having the concordant dysregulated direction in two lists of genes by chance.

2.6 | Functional enrichment analysis

The gene categories for functional enrichment analysis were performed on the Kyoto Encyclopedia of Genes and Genomes.[45] The hypergeometric distribution model[46] was applied to determine the significance of biological pathways enriched by genes of interest. The Benjamini and Hochberg procedure[47] was used to estimate the FDR. Statistical analysis was carried out with the R software package version 3.5.1.

3 | RESULTS

3.1 | Development of the qualitative REOs-based signature

The general workflow of this study is described in Figure 1. Total, 131 FF PCa samples measured by the GPL10264 platform (Table 1), denoted as PC131, were used as the training data set. Using the univariate Cox proportional-hazards regression model with FDR < 20%, we found 80 genes with expression levels significantly correlated with the BFS of PCa patients after radical prostatectomy. A total of 3160 gene pairs consisting of every two of the 80 prognosis-associated genes were constructed and each gene pair classified all samples into two subgroups according to its REO in each sample. Using the univariate Cox proportional-hazards regression model with FDR < 20%, 1205 prognosis-associated gene pairs were identified and sorted in descending order according to their C-index values. Then, based on a forward selection method (see Section 2), 74 gene pairs with the highest C-index (C-index = 0.87) were chosen as the final prognostic signature, denoted as 74-GPS (Table 2). Patients were classified into the high-risk group when at least 37 of 74 gene pairs suggested that this patient was at high risk; otherwise, the low-risk group. According to this decision rule, samples in the training data set were stratified into two subgroups: 108 samples in the low-risk group and 23 samples in the high-risk group, and the BFS of the former group were significantly better than the latter group (HR = 63.23, 95% CI: 18.56-215.40, $P < 2.2 \times 10^{-16}$, C-index = 0.87, Figure 2A). A multivariate Cox regression analysis revealed that the 74-GPS still displayed significant correlations with patients' BFS in the training data set if the clinical factors of Gleason grade group, surgical margins, preoperative PSA and pathological stage were considered, as shown in Figure 3.

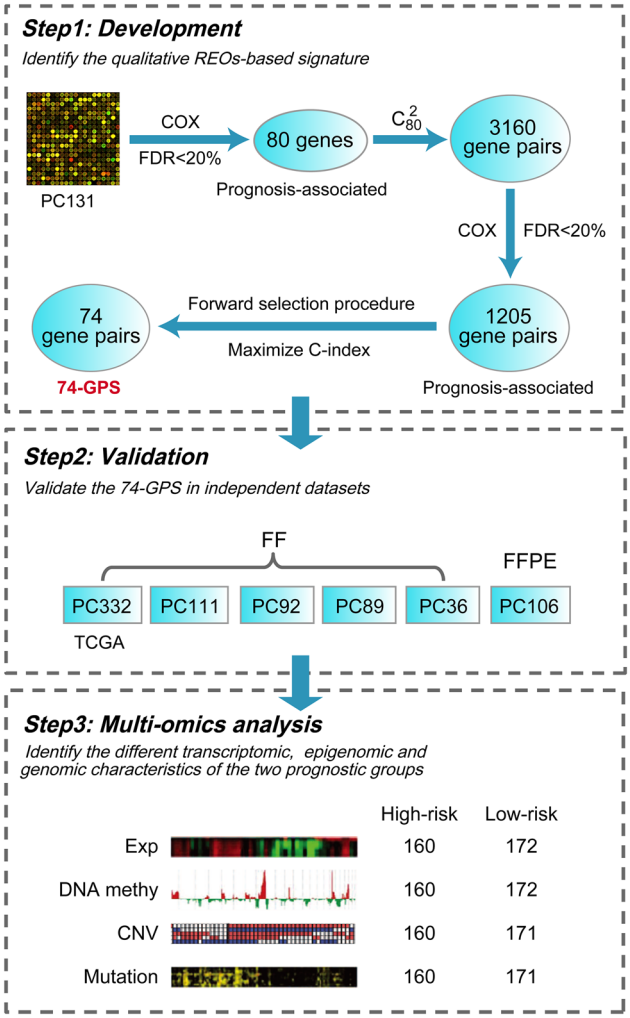

FIGURE 1 Overview of the workflow used in this study. The workflow includes three major steps: the development of the REOs-based signature in the training datasets (Step 1), the validation of the signature in the six independent validation datasets (Step 2), and the multiomics characteristics analyses of the two prognostic groups (Step 3). CNV, copy number variation; DNA methy, DNA methylation; Exp, expression; REOs, relative expression orderings [Color figure can be viewed at wileyonlinelibrary.com]

3.2 | Validation of the qualitative REOs-based signature

The first validation data set included 332 FF samples from TCGA, denoted as PC332. The 172 patients predicted to be at the low-risk recurrence group had a significantly better BFS than the 160 patients predicted to be at the high-risk group (HR = 2.02, 95% CI: 1.06-3.83, $P = 2.78 \times 10^{-2}$, C-index = 0.60; Figure 2B). The second validation data set included 111 FF samples measured by the GPL10558 platform, denoted as PC111. The signature classified 98

TABLE 2 The composition of the 74-GPS

Pair 1-25	Gene A	Gene B	Pair 26-50	Gene A	Gene B	Pair 51-74	Gene A	Gene B
pair1	*ENG*	*CEBPD*	pair26	*ASPN*	*ZNF622*	pair51	*COL5A2*	*HELB*
pair2	*INHBA*	*CFDP1*	pair27	*INHBA*	*ITGA11*	pair52	*RELN*	*CTHRC1*
pair3	*ZHX3*	*TACC2*	pair28	*ASPN*	*OLFML2B*	pair53	*BGN*	*TACC2*
pair4	*OLFML2B*	*HELB*	pair29	*ENG*	*TJP2*	pair54	*PPP2R2C*	*TACC2*
pair5	*COL1A1*	*HSPA1B*	pair30	*ZHX3*	*CCNL1*	pair55	*FOLH1B*	*TACC2*
pair6	*NOTCH3*	*CEBPD*	pair31	*LTBP2*	*ZNF334*	pair56	*THBS2*	*CEBPD*
pair7	*PPP2R2C*	*CFDP1*	pair32	*RELN*	*CLEC14A*	pair57	*BGN*	*FZD5*
pair8	*COL3A1*	*ZFP36*	pair33	*COL8A1*	*HELB*	pair58	*POSTN*	*FZD5*
pair9	*COL1A1*	*NXF1*	pair34	*TACC2*	*ZNF532*	pair59	*COMP*	*DLL4*
pair10	*THBS2*	*CCNL1*	pair35	*FOLH1B*	*ZNF532*	pair60	*SFRP4*	*JUNB*
pair11	*ZHX3*	*TEP1*	pair36	*LTBP2*	*CCNL1*	pair61	*COL8A1*	*HOPX*
pair12	*NIPA1*	*ZNF532*	pair37	*COL3A1*	*TJP2*	pair62	*ESM1*	*HOPX*
pair13	*COL3A1*	*TACC2*	pair38	*CLSTN2*	*FZD5*	pair63	*LTBP2*	*CFDP1*
pair14	*XPO6*	*NXF1*	pair39	*COL3A1*	*FZD5*	pair64	*CLSTN2*	*ZFP36*
pair15	*HOPX*	*HELB*	pair40	*CLSTN2*	*CCNL1*	pair65	*FOLH1*	*CEBPD*
pair16	*CDH13*	*HELB*	pair41	*TCF19*	*HELB*	pair66	*FAP*	*LAMP5*
pair17	*POSTN*	*SLC25A17*	pair42	*OR,2T2*	*CTHRC1*	pair67	*CCNL1*	*FZD5*
pair18	*NIPA1*	*CCNL1*	pair43	*SFRP4*	*HOPX*	pair68	*CXCL14*	*TACC2*
pair19	*ASPN*	*NOX4*	pair44	*NOTCH3*	*XPO6*	pair69	*CTHRC1*	*DLL4*
pair20	*FOLH1B*	*HSPA1B*	pair45	*PPP2R2C*	*ITGA11*	pair70	*PYDC2*	*MAB21L3*
pair21	*COL3A1*	*XPO6*	pair46	*OR,2T11*	*OLFML2B*	pair71	*COL10A1*	*ABCC11*
pair22	*HOXC4*	*HELB*	pair47	*THY1*	*ITGA11*	pair72	*CDH13*	*ZNF334*
pair23	*COL1A1*	*CCNL1*	pair48	*CXCL14*	*CCNL1*	pair73	*CCNL1*	*CEBPD*
pair24	*COMP*	*TJP2*	pair49	*DLL4*	*HELB*	pair74	*CPS1*	*COL5A2*
pair25	*COL3A1*	*ENG*	pair50	*NIPA1*	*ZFP36*			

Note: Gene pair votes for high-risk when Gene A has a higher expression level than Gene B in a sample.

and 13 samples into the low-risk and high-risk groups, respectively while the BFS of the former were significantly better than the latter (HR = 4.69, 95% CI: 1.63-13.51, $P = 1.15 \times 10^{-2}$, C-index = 0.61; Figure 2C). The signature was also verified in the other three validation datasets with 92, 89, and 36 FF samples, respectively. Each group of patients at the low-risk had significantly longer BFS than the group of patients at the high-risk in all datasets (Figure 2D-F). Notably, 106 FFPE samples in the data set GSE54460 were successfully stratified into two different prognostic groups: 91 patients at the low-risk group had a significantly better BFS than 15 patients at the high-risk group (HR = 2.68, 95% CI: 1.40-5.14, $P = 6.82 \times 10^{-3}$, C-index = 0.57; Figure 2G). As we know, FFPE samples always suffer RNA degrade during the process of preparation and storage, which hampers the clinical application of quantitative transcriptional signatures.[27,48]

The multivariate Cox regression analysis was also performed in the validation datasets. The results showed that the signature remained significantly associated with patients' BFS in the datasets PC332, PC111, PC92, and PC89 after adjusting the available clinic pathological factors. The detailed information was shown in Figure 4 and Table S1.

3.3 | Distinct transcriptional and functional characteristics of the two prognostic groups

With 10% FDR control, 177 DEGs and 1250 DEGs were identified by RankCompV2 between the high- and low-risk prognostic groups of the datasets PC131 and PC332, respectively. These two lists of DEGs shared 84 genes, of which 83 genes showed the same dysregulated directions in the high-risk group compared with the low-risk group, with a direction concordance score of 98.81% which was unlikely observed by chance (binomial distribution test, $P < 2.2 \times 10^{-16}$, see Section 2). Besides, the direction concordance scores between every two of the DEGs lists detected from the seven datasets were all unlikely happened by chance (see Table S2). These results suggested that the distinct transcriptional characteristics of the two prognostic groups were highly reproducible in the independent datasets.

With FDR < 10%, functional enrichment analysis for the 1250 DEGs identified from TCGA samples in the data set PC332 revealed that the genes upregulated in the high-risk group were significantly enriched in pathways associated with cell proliferation, such as the PI3K-Akt signaling pathway and the TGF-beta signaling pathway, whereas the downregulated genes were significantly enriched in

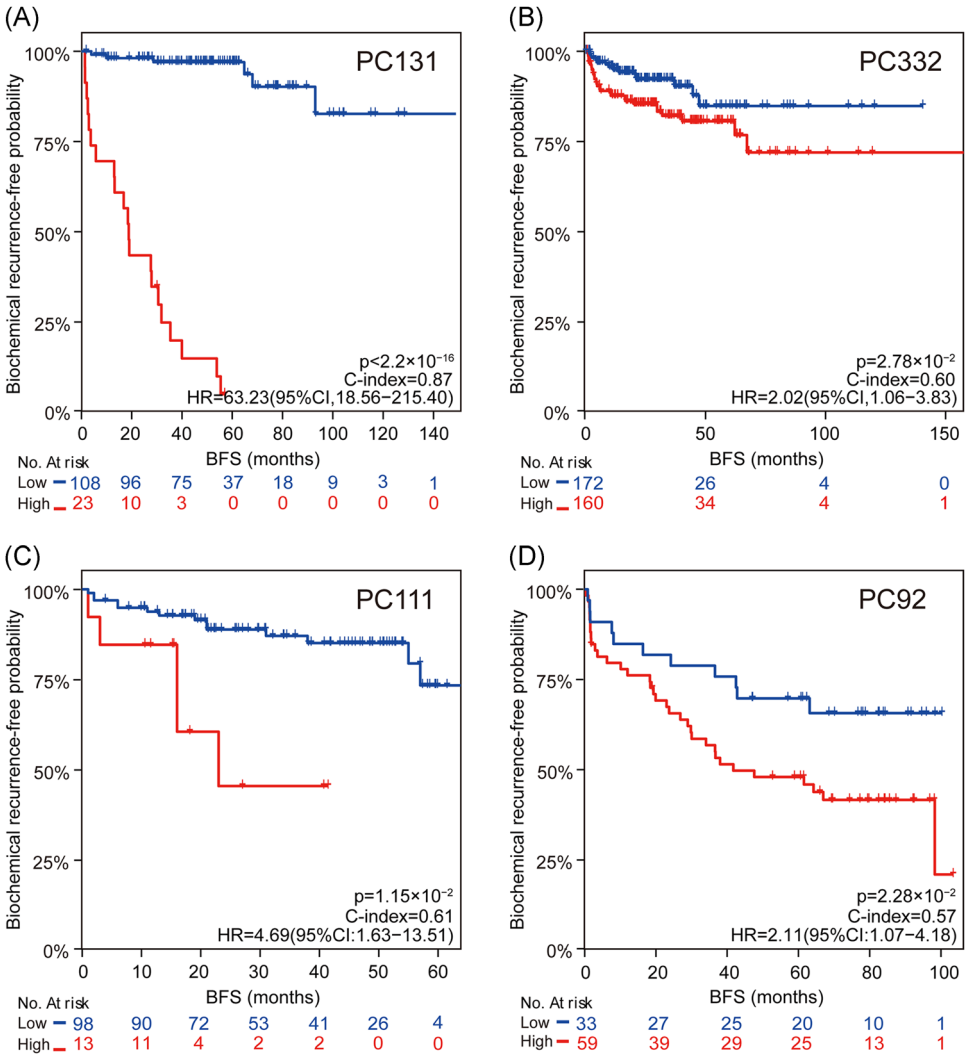

FIGURE 2 The Kaplan-Meier curves of biochemical recurrence-free survival for the training and validation datasets. The Kaplan-Meier curves of biochemical recurrence-free survival for the training data set PC131 (A) and the six validation datasets PC332 (B), PC111 (C), PC92 (D), PC89 (E), PC36 (F), and PC106 (G). A sample was assigned into the high-risk group (red lines) when at least 37 gene pairs of the 74-GPS voted for high-risk; otherwise, the low risk group (blue lines). GPS, gene pairs [Color figure can be viewed at wileyonlinelibrary.com]

metabolic pathways, such as fatty acid degradation pathway and glutathione metabolism pathway (hypergeometric distribution model, FDR < 10%, Table S3). These results indicated that the tumor cells in the high-risk patients might grow faster than that in the low-risk patients and experience dysregulated metabolism, which led to poor prognosis of PCa patients.[49,50]

3.4 | Distinct epigenomic characteristics of the two prognostic groups

In the TCGA data set PC332, 160 samples and 172 samples with DNA methylation data were classified into the high-risk prognostic group and the low-risk prognostic group by the 74-GPS, respectively. Using the Wilcoxon rank-sum test with FDR < 1%, 1631 hypermethylated and 624 hypomethylated genes were identified from the high-risk prognostic group compared with the low-risk prognostic group, respectively. There were 12.94% of 1631 hypermethylated genes overlapped with the 1250 DEGs between the two different prognostic groups. The direction concordance score of hypermethylation with downregulation was 94.31%, which was extremely unlikely happened due to chance (binomial distribution test, $P < 2.2 \times 10^{-16}$). Similarly, 11.86% of 624 hypomethylated genes were overlapped with DEGs between the high-risk prognostic groups and low-risk prognostic groups. The direction concordance score of hypomethylation with upregulation was 93.24%, which was also extremely unlikely happened due to chance (binomial distribution test, $P = 8.88 \times 10^{-16}$).

An additional 160 samples without the recurrence information in the TCGA data portal, denoted as PC160, were used to confirm the epigenomic characteristics of the two prognostic groups. In data set

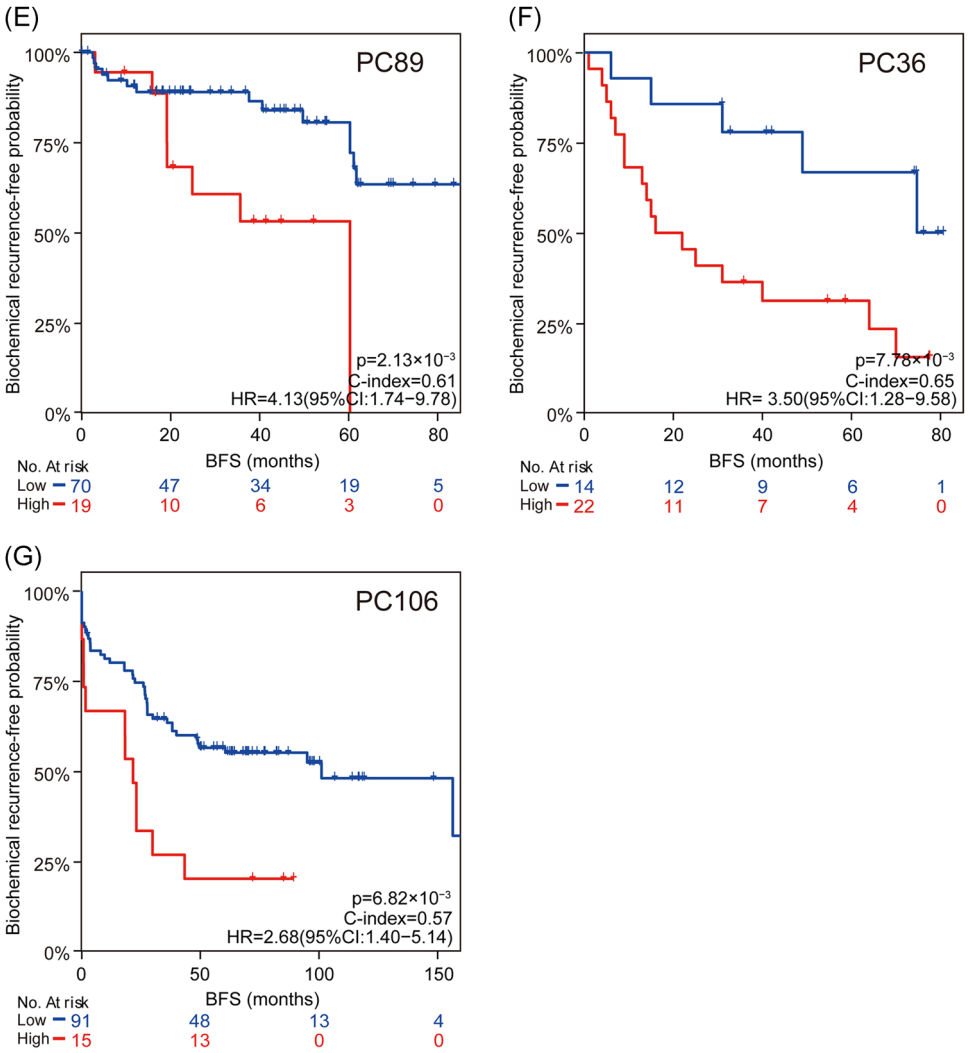

FIGURE 2 Continued

PC160, 98 samples and 62 samples with DNA methylation data were classified into the high-risk prognostic group and the low-risk prognostic group by the 74-GPS. With 10% FDR control, 588 DEGs were identified by RankCompV2 between the high- and low-risk prognostic groups. The direction concordance score of DEGs from data set PC160 and data set PC332 was 100% (425/425). Using the Wilcoxon rank-sum test with FDR < 10%, 542 hypermethylated and 237 hypomethylated genes were identified from the high-risk

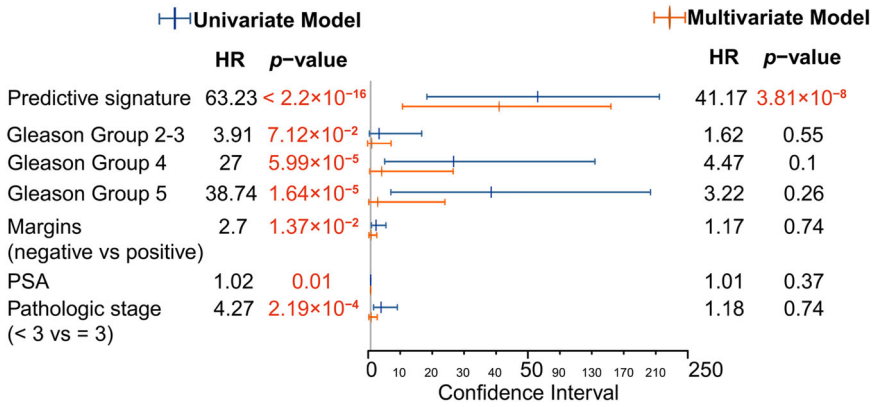

FIGURE 3 Univariate and multivariate Cox regression analyses for the 74-GPS in the training data set. The forest plot of univariate (blue lines) and multivariate (orange lines) Cox regression analysis of the predictive signature and available prognostic factors in the training data set PC131. Red color indicates significant *P* values. *P* < .1. GPS, gene pairs [Color figure can be viewed at wileyonlinelibrary.com]

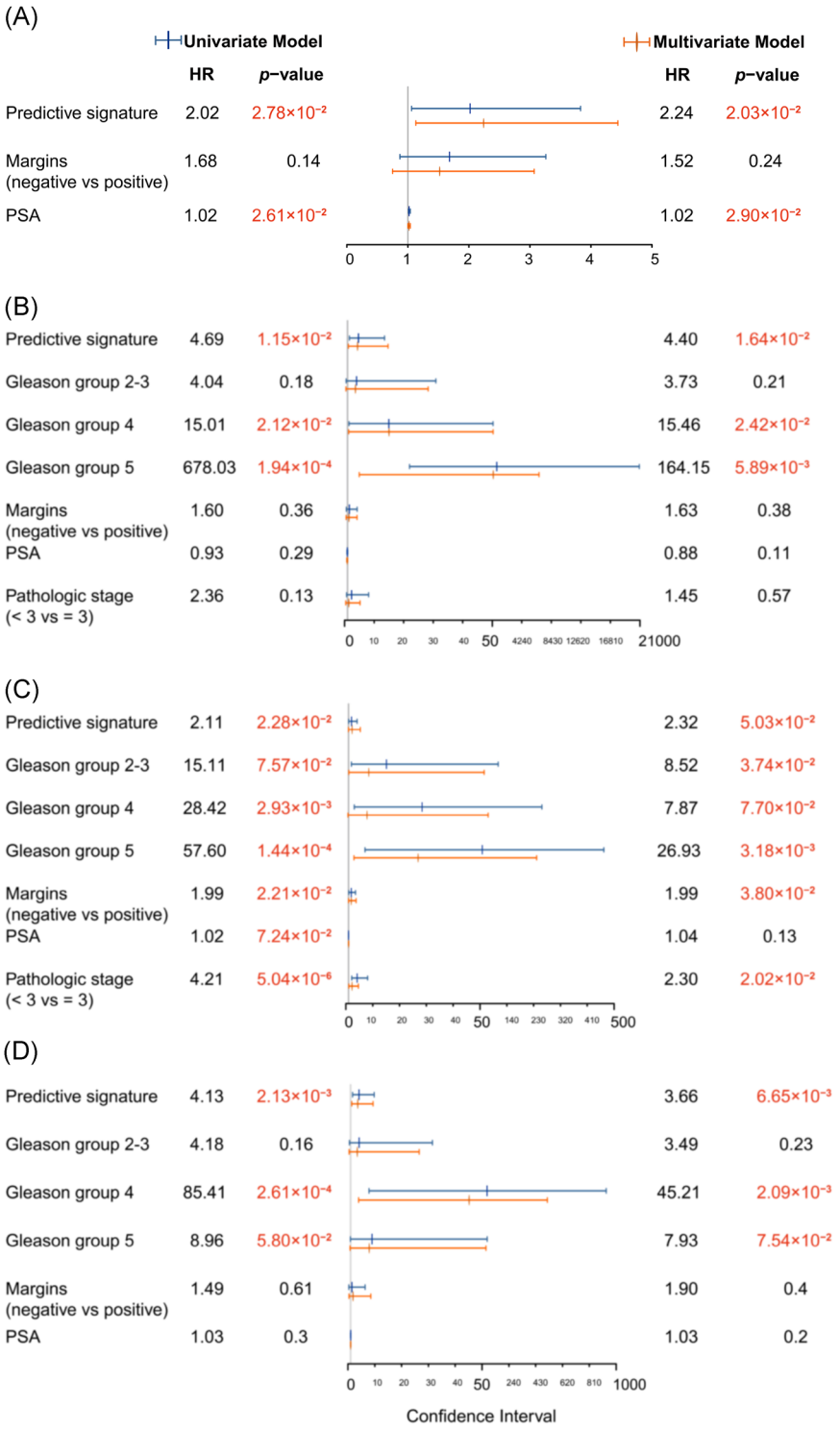

FIGURE 4 Univariate and multivariate Cox regression analyses for the 74-GPS in the validation datasets. The forest plot of univariate (blue lines) and multivariate (orange lines) Cox regression analyses of the predictive signature and available prognostic factors in the validation datasets PC332 (A), PC111 (B), PC92 (C), and PC89 (D). Red color indicates significant *P* values. $P < .1$. GPS, gene pairs [Color figure can be viewed at wileyonlinelibrary.com]

prognostic group compared with the low-risk prognostic group, respectively. The direction concordance score of DMGs from data set PC160 and data set PC332 was 97.98% (339/346). Moreover, the direction concordance scores of hypermethylation with downregulation and hypomethylation with upregulation in data set PC160 were 95.83% (23/24) and 97.50% (39/40), respectively. All the direction concordance scores were extremely unlikely happened due to chance, as shown in Table S4 and S5. The consistent and reproducible results observed in datasets PC332 and PC160 implied that the epigenetic alterations may cause reproducibly transcriptional alterations between different prognostic groups.

3.5 | Distinct genomic characteristics of the two prognostic groups

The 331 samples with copy number alteration data in the data set PC332 were divided into 160 high-risk samples and 171 low-risk

samples, respectively. A total of 10,342 genes were with significantly higher copy number alteration frequencies in the high-risk group than in the low-risk group (Fisher exact test, FDR < 5%). Then 6.48% of 2378 genes frequently amplified in the high-risk group were shared in the DEGs list, of which 68.18% were upregulated in the high-risk group. This was unlikely happened due to chance (binomial distribution test, $P = 3.75 \times 10^{-6}$). Moreover, Among the 7964 genes frequently deleted in the high-risk group, 457 genes were shared in the DEGs list, of which 61.49% were downregulated in the high-risk group. This was also unlikely happened due to chance (binomial distribution test, $P = 5.16 \times 10^{-7}$).

The 109 samples with copy number variation data in the data set PC131 were divided into 24 high-risk samples and 85 low-risk samples by the 74-GPS, respectively. With 5% *p*-value control, 4020 DEGs were identified by Student`s *t*-test between the high- and low-risk prognostic groups with a direction concordance score of 97.08% (333/343) in the data set PC332. A total of 2957 genes had significantly higher copy number alteration frequencies in the high-risk group than in the low-risk group (Fisher exact test, $P < 5\%$), with a direction concordance score of 94.40% (1483/1571) in the datasets PC332. And the direction concordance scores of amplification with upregulation and deletion with downregulation in the high-risk group of the data set PC131 were 74.47% (70/94) and 61.80% (406/657), respectively. All the direction concordance scores were unlikely happened due to chance, as shown in Table S4 and S6.

In the data set PC160, 98 samples and 62 samples with copy number variation data were classified into the high-risk prognostic group and the low-risk prognostic group, respectively. A total of 4893 genes had significantly higher copy number alteration frequencies in the high-risk group than in the low-risk group (Fisher exact test, $P < 5\%$), with a direction concordance score of 74.47% (2138/2871) in the datasets PC332. And the direction concordance score of amplification with upregulation in the high-risk group was 82.26% (51/62). All the direction concordance scores were also unlikely happened due to chance (shown in Table S4 and S6).

The results observed in the datasets PC332, PC131, and PC160 implied that these copy number alterations, especially amplification, may cause reproducibly transcriptional alterations between different prognostic groups.

Using Fisher's exact test with $P < .1$, 84 genes whose mutation frequencies tended to be different were detected between the 160 high-risk samples and the 171 low-risk samples with somatic mutation data in the data set PC332 (Table S7). Impressively, all of the 84 genes had higher mutation frequencies in the high-risk group than in the low-risk group, which was unlikely to happen due to chance (binomial distribution test, $P < 2.2 \times 10^{-16}$). In the data set PC160, 12 genes whose mutation frequencies tended to be different were detected between the 97 high-risk and the 62 low-risk samples with somatic mutation data (Fisher's exact test, $P < .1$). In both the data set PC160 and PC332, *TP53* was with significantly higher mutation frequencies in the high-risk group than in the low-risk group. It has been reported that *TP53* mutation is correlated with metastasis and poor prognosis of PCa patients.[51–53] Functional enrichment analysis showed that these 84 mutation genes were significantly enriched in focal adhesion and PI3K-Akt signaling pathways (hypergeometric distribution model, FDR < 10%), suggesting that mutation-induced alternation of genes in these pathways might lead a poor outcome of PCa patients.

4 | DISCUSSION

In this study, we developed a qualitative transcriptional signature, 74-GPS, to predict the recurrence risk of PCa patients after radical prostatectomy, which was validated in six independent datasets produced by different platforms, including a total of 660 FF and 106 FFPE samples. The further multiomics data analyses showed that the distinct transcriptomic, epigenetic, and genomic landscapes between the high-risk and low-risk groups might be helpful to understand the mechanisms of different prognoses and prescribe more specific and proper treatments for PCa patients. Consistent with our previous study,[34] the qualitative REOs-based prognostic signature is highly robust against experimental batch effects and differences in probe designs of different platforms. Besides, the signature can be readily applied at an individualized level without data normalization, which is more reliable and practical than quantitative signatures for risk prediction.[54]

At present, most of the clinical tissue samples are fixed in FFPE blocks,[55–57] and stored in hospitals and tissue banks, which is a huge and precious resource for clinical research.[58] Nevertheless, FFPE samples are generally considered unreliable for gene expression analysis because of RNA degradation during preparation and storage. As shown in our previous study,[27] the expression measurements of thousands of genes had at least two-fold change in FFPE samples compared with paired FF samples. Therefore, quantitative signatures based on gene expression measurements of FFPE (or FF) samples could not be applied to FF (or FFPE) samples directly. In contrast, as demonstrated in our previous study[27] and confirmed in this study, most of the REOs of gene pairs in FFPE samples were insensitive to partial RNA degradation, which makes it possible to be easily applied to both FF and FFPE samples.

One limitation of our study was that all samples for the development and validation of the signature were from the public databases. We noticed that compared with preoperative PSA and surgical margins, the signature lost significance in the datasets PC36 and PC106, which might be attributed to the inherent limitations in the public domain data available, such as the small sample size of data set PC36 or the poor quality of gene expression measurements for FFPE samples in data set PC106. Although the multivariate Cox regression analysis of samples integrated from datasets PC111, PC92, PC36, and PC106 with the common available clinical-pathological factors showed that the signature remained significantly associated with patients' BFS (shown in Table S1 and Figure S1), it is necessary to collect an additional data set of independent samples in our future work to validate our signature. For the sake of more reliable prediction

under some circumstances, the 74-GPS could be combined with preoperative PSA and surgical margins to predict the biochemical recurrence for PCa patients.

The multiomics analysis of TCGA samples played an essential role in uncovering the underlying molecular mechanisms of determining different prognoses of PCa patients after radical prostatectomy. For example, gene *PDGFRB* in the PI3K-Akt signaling pathway, which was upregulated with concordant hypomethylation in the high-risk group, can regulate cell growth, division, and migration[59,60] and is correlated with bone metastases and biochemical recurrence of PCa patients.[61,62] In addition, gene *THBS1* in the TGF-beta signaling pathway, which was upregulated with consistent hypomethylation alteration in the high-risk group, has been reported to be positively associated with the invasion of PCa and the recurrence of PCa patients after radical prostatectomy.[63] These results provided evidence that the tumor cells of high-risk patients own faster growth and stronger migration abilities, which result in poorer prognoses.

According to the National Comprehensive Cancer Network (NCCN) guidelines for prostate cancer patients after radical prostatectomy,[64] PSA measurements should be performed every 6 to 12 months and a digital rectal examination (DRE) is recommended annually for the first 5 years. For PCa patients assigned as high risk by the signature, we suggest that more close follow-ups, such as PSA testing every 3 months and DRE every 6 months for the first 5 years, maybe better to detect disease progression timely. This study may be helpful to guide management and improve prognoses for PCa patients after radical prostatectomy.

5 | CONCLUSION

In conclusion, the qualitative REO-based 74-GPS is a robust individual-level prognostic signature for predicting the BFS of postsurgical PCa patients from different hospitals equipped with different platforms. The PCa patients who identified with a high risk of biochemical recurrence by the signature should have timely treatments or close follow-ups.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant Nos: 81602738 and 81872396), young and middle-aged backbone training project in the health system of Fujian province (Grant No. 2017-ZQN-56), outstanding youth scientific research personnel training program in Fujian Province University (Grant No. 2017B018), and the Joint Scientific and Technology Innovation Fund of Fujian Province (Grant No: 2018Y9065).

AUTHOR CONTRIBUTIONS

LA and ZG designed and supervised the research study; XL and LA performed the research; XL, JHZ, FLJ, YTG, and YDS performed the data analysis; XL, HYH, and LA wrote the R codes; XL and LA drafted the manuscript; LA and ZG revised the manuscript; XL and YTG interpreted the function annotations; XL and HYH drew the figures. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare that they have no conflict of interests.

ORCID

Lu Ao http://orcid.org/0000-0001-7378-4967

REFERENCES

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA Cancer J Clin*. 2015;65(2):87-108.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424.
3. Chen W, Zheng R, Baade PD, et al. Cancer statistics in China, 2015. *CA Cancer J Clin*. 2016;66(2):115-132.
4. Han M, Partin AW, Zahurak M, Piantadosi S, Epstein JI, Walsh PC. Biochemical (prostate specific antigen) recurrence probability following radical prostatectomy for clinically localized prostate cancer. *J Urol*. 2003;169(2):517-523.
5. Ward JF, Moul JW. Rising prostate-specific antigen after primary prostate cancer therapy. *Nat Clin Pract Urol*. 2005;2(4):174-182.
6. Mottet N, Bellmunt J, Bolla M, et al. EAU-ESTRO-SIOG guidelines on prostate cancer. Part 1: screening, diagnosis, and local treatment with curative intent. *Eur Urol*. 2017;71(4):618-629.
7. Sandblom G, Ladjevardi S, Garmo H, Varenhorst E. The impact of prostate-specific antigen level at diagnosis on the relative survival of 28,531 men with localized carcinoma of the prostate. *Cancer*. 2008;112(4):813-819.
8. Tollefson MK, Blute ML, Rangel LJ, Bergstralh EJ, Boorjian SA, Karnes RJ. The effect of Gleason score on the predictive value of prostate-specific antigen doubling time. *BJU Int*. 2010;105(10):1381-1385.
9. Kattan MW, Wheeler TM, Scardino PT. Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. *J Clin Oncol*. 1999;17(5):1499-1507.
10. Hull G, Rabbani F, Abbas F, Wheeler T, Kattan M, Scardino P. Cancer control with radical prostatectomy alone In 1,000 consecutive patients. *J Urol*. 2002;167(2):528-534.
11. D'Amico AV, Whittington R, Malkowicz SB, et al. Combination of the preoperative PSA level, biopsy Gleason score, percentage of positive biopsies, and MRI T-stage to predict early PSA failure in men with clinically localized prostate cancer. *Urology*. 2000;55(4):572-577.
12. Tsivian M, Sun L, Mouraviev V, et al. Changes in Gleason score grading and their effect in predicting outcome after radical prostatectomy. *Urology*. 2009;74(5):1090-1093.
13. Budäus L, Isbarn H, Eichelberg C, et al. Biochemical recurrence after radical prostatectomy: multiplicative interaction between surgical margin status and pathological stage. *J Urol*. 2010;184(4):1341-1346.
14. Bielinsky A-K, Nakagawa T, Kollmeyer TM, et al. A tissue biomarker panel predicting systemic progression after PSA recurrence post-definitive prostate cancer therapy. *PLoS One*. 2008;3(5):e2318.
15. Chandran UR, Ma C, Dhir R, et al. Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC Cancer*. 2007;7:64.

16. Kim PJ, Park JY, Kim HG, Cho YM, Go H. Dishevelled segment polarity protein 3 (DVL3): a novel and easily applicable recurrence predictor in localised prostate adenocarcinoma. *BJU Int*. 2017;120(3):343-350.
17. Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*. 2010;11(10):733-739.
18. Lazar C, Meganck S, Taminau J, et al. Batch effect removal methods for microarray gene expression data integration: a survey. *Brief Bioinform*. 2013;14(4):469-490.
19. Ferte C, Trister AD, Huang E, et al. Impact of bioinformatic procedures in the development and translation of high-throughput molecular classifiers in oncology. *Clin Cancer Res*. 2013;19(16):4315-4325.
20. Eddy JA, Sung J, Geman D, Price ND. Relative expression analysis for molecular cancer diagnosis and prognosis. *Technol Cancer Res Treat*. 2010;9(2):149-159.
21. Patil P, Bachant-Winner PO, Haibe-Kains B, Leek JT. Test set bias affects reproducibility of gene signatures. *Bioinformatics*. 2015;31(14):2318-2323.
22. Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mRNA comparisons. *Stat Appl Genet Mol Biol*. 2004;3:1-19. Article19.
23. Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*. 2005;21(20):3896-3904.
24. Marchionni L, Afsari B, Geman D, Leek JT. A simple and reproducible breast cancer prognostic test. *BMC Genomics*. 2013;14:336.
25. Afsari B, Braga-Neto UM, Geman D. Rank discriminants for predicting phenotypes from RNA expression. *Ann Appl Stat*. 2014;8(3):1469-1491.
26. Cheng J, Guo Y, Gao Q, et al. Circumvent the uncertainty in the applications of transcriptional signatures to tumor tissues sampled from different tumor sites. *Oncotarget*. 2017;8(18):30265-30275.
27. Chen R, Guan Q, Cheng J, et al. Robust transcriptional tumor signatures applicable to both formalin-fixed paraffin-embedded and fresh-frozen samples. *Oncotarget*. 2017;8(4):6652-6662.
28. Ao L, Zhang Z, Guan Q, et al. A qualitative signature for early diagnosis of hepatocellular carcinoma based on relative expression orderings. *Liver Int*. 2018;38(10):1812-1819.
29. Liu H, Li Y, He J, et al. Robust transcriptional signatures for low-input RNA samples based on relative expression orderings. *BMC Genomics*. 2017;18(1):913.
30. Cai H, Li X, Li J, et al. Tamoxifen therapy benefit predictive signature coupled with prognostic signature of post-operative recurrent risk for early stage ER+breast cancer. *Oncotarget*. 2015;6(42):44593-44608.
31. Zhang L, Hao C, Shen X, et al. Rank-based predictors for response and prognosis of neoadjuvant taxane-anthracycline-based chemotherapy in breast cancer. *Breast Cancer Res Treat*. 2013;139(2):361-369.
32. Zhao W, Chen B, Guo X, et al. A rank-based transcriptional signature for predicting relapse risk of stage II colorectal cancer identified with proper data sources. *Oncotarget*. 2016;7(14):19060-19071.
33. Li X, Cai H, Zheng W, et al. An individualized prognostic signature for gastric cancer patients treated with 5-Fluorouracil-based chemotherapy and distinct multi-omics characteristics of prognostic groups. *Oncotarget*. 2016;7(8):8743-8755.
34. Ao L, Song X, Li X, et al. An individualized prognostic signature and multiomics distinction for early stage hepatocellular carcinoma patients with surgical resection. *Oncotarget*. 2016;7(17):24097-24110.
35. Qi L, Chen L, Li Y, et al. Critical limitations of prognostic signatures based on risk scores summarized from gene expression levels: a case study for resected stage I non-small-cell lung cancer. *Brief Bioinform*. 2016;17(2):233-242.
36. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207-210.
37. Irizarry RA. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4(2):249-264.
38. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*. 2013;6(269):pl1.
39. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2012;2(5):401-404.
40. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361-387.
41. Schweder T, Spjøtvoll E. A class of rank test procedures for censored survival data. *Biometrika*. 1982;69(3):553-566.
42. Storey JD. A direct approach to false discovery rates. *J Royal Stat Soc*. 2002;64(3):479-498.
43. Cai H, Li X, Li J, et al. Identifying differentially expressed genes from cross-site integrated data based on relative expression orderings. *Int J Biol Sci*. 2018;14(8):892-900.
44. Bahn AK. Application of binomial distribution to medicine: comparison of one sample proportion to an expected proportion (for small samples). Evaluation of a new treatment. Evaluation of a risk factor. *J Am Med Womens Assoc*. 1969;24(12):957-966.
45. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28(1):27-30.
46. Hong G, Zhang W, Li H, Shen X, Guo Z. Separate enrichment analysis of pathways for up- and downregulated genes. *J R Soc Interface*. 2014;11(92):20130950.
47. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc*. 1995;57:289-300.
48. Waldron L, Ogino S, Hoshida Y, et al. Expression profiling of archival tumors for long-term health studies. *Clin Cancer Res*. 2012;18(22):6136-6146.
49. Tang KW, Alaei-Mahabadi B, Samuelsson T, Lindh M, Larsson E. The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat Commun*. 2013;4:2513.
50. Zadra G, Photopoulos C, Loda M. The fat side of prostate cancer. *Biochim Biophys Acta*. 2013;1831(10):1518-1532.
51. Sirohi D, Devine P, Grenert JP, van Ziffle J, Simko JP, Stohr BA. TP53 structural variants in metastatic prostatic carcinoma. *PLoS One*. 2019;14(6):e0218618.
52. Ecke TH, Schlechte HH, Schiemenz K, et al. TP53 gene mutations in prostate cancer progression. *Anticancer Res*. 2010;30(5):1579-1586.
53. Hong MKH, Macintyre G, Wedge DC, et al. Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. *Nat Commun*. 2015;6:6605.
54. Qi L, Li T, Shi G, et al. An individualized gene expression signature for prediction of lung adenocarcinoma metastases. *Mol Oncol*. 2017;11(11):1630-1645.
55. Abdullah-Sayani A, Bueno-de-Mesquita JM, van de Vijver MJ. Technology Insight: tuning into the genetic orchestra using microarrays--limitations of DNA microarrays in clinical practice. *Nat Clin Pract Oncol*. 2006;3(9):501-516.
56. von Ahlfen S, Missel A, Bendrat K, Schlumpberger M. Determinants of RNA quality from FFPE samples. *PLoS One*. 2007;2(12):e1261.
57. Thomas M, Poignee-Heger M, Weisser M, Wessner S, Belousov A. An optimized workflow for improved gene expression profiling for formalin-fixed, paraffin-embedded tumor samples. *J Clin Bioinform*. 2013;3(1):10.
58. Blow N. Tissue preparation: tissue issues. *Nature*. 2007;448(7156):959-963.

59. Appiah-Kubi K, Wang Y, Qian H, et al. Platelet-derived growth factor receptor/platelet-derived growth factor (PDGFR/PDGF) system is a prognostic and treatment response biomarker with multifarious therapeutic targets in cancers. *Tumour Biol*. 2016;37(8):10053-10066.
60. Andrae J, Gallini R, Betsholtz C. Role of platelet-derived growth factors in physiology and medicine. *Genes Dev*. 2008;22(10):1276-1312.
61. Dolloff NG, Shulby SS, Nelson AV, et al. Bone-metastatic potential of human prostate cancer cells correlates with Akt/PKB activation by alpha platelet-derived growth factor receptor. *Oncogene*. 2005;24(45):6848-6854.
62. Nordby Y, Richardsen E, Rakaee M, et al. High expression of PDGFR-beta in prostate cancer stroma is independently associated with clinical and biochemical prostate cancer recurrence. *Sci Rep*. 2017;7:43378.
63. Firlej V, Mathieu JRR, Gilbert C, et al. Thrombospondin-1 triggers cell migration and development of advanced prostate tumors. *Cancer Res*. 2011;71(24):7649-7658.
64. Mohler JL, Antonarakis ES, Armstrong AJ, et al. Prostate cancer, Version 2.2019, NCCN clinical practice guidelines in oncology. *J Natl Compr Canc Netw*. 2019;17(5):479-505.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Li X, Huang H, Zhang J, et al. A qualitative transcriptional signature for predicting the biochemical recurrence risk of prostate cancer patients after radical prostatectomy. *The Prostate*. 2020;80:376–387. https://doi.org/10.1002/pros.23952